Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks

Anthony Rousseau, Paul Deléglise, Yannick Estève
Laboratoire Informatique de l'Université du Maine (LIUM)
University of Le Mans, France
firstname.lastname@lium.univ-lemans.fr

Abstract

In this paper, we present improvements made to the TED-LIUM corpus we released in 2012. These enhancements fall into two categories. First, we describe how we filtered publicly available monolingual data and used it to estimate well-suited language models (LMs), using open-source tools. Then, we describe the process of selection we applied to new acoustic data from TED talks, providing additions to our previously released corpus. Finally, we report some experiments we made around these improvements.

Keywords: Corpus, Speech Recognition, Language Modeling

1. Introduction

Back in May 2012, we presented at LREC a new corpus dedicated to Automatic Speech Recognition we developed to participate to the IWSLT2011 campaign (Federico et al., 2011) on the Spoken Language Translation task (spoken English to written French). This corpus is composed with segments of public talks extracted from the TED website[1] in order to be as close as possible to the content of the evaluation sets from the aforementioned task. We believe this helped us to propose the best-ranked system during this campaign, while the great diversity of speakers also ensure a reasonable generalization for other tasks.

In this paper, we propose enhancements we made to this corpus in a new release which will be made publicly available later this year. These improvements are of two kinds:

- addition of monolingual text data aimed at language modeling, filtered with data selection techniques, well-suited for decoding tasks using our TED-LIUM corpus and, in our experiments, leading to interesting WER reductions,
- addition of new acoustic data extracted from TED talks, along with corresponding automatically aligned transcripts and an updated training dictionary, also leading to a decrease in the word error rate of our system.

The remainder of this paper is organized as follows: first, in section 2., we briefly summarize the original release of the TED-LIUM corpus. Then, in section 3., we describe the data selection experiments we made in order to improve the language modeling part of our system using only open-source tools and publicly available data. Finally, in section 4., we present the segment selection process we used for new TED Talks inclusion and the characteristics of this new acoustic data for our corpus. We also discuss the improvements obtained using this new data in our system.

[1] http://www.ted.com

2. The TED-LIUM corpus

The TED-LIUM corpus was initially released in May 2012, during the LREC'12 conference in Istanbul, Turkey (Rousseau et al., 2012). It was developed within the context of the LIUM's participation to the IWSLT 2011 evaluation campaign. All its raw data (acoustic signals and closed captions) was extracted from the TED website, and automatic transcriptions obtained from decoding the acoustic signals were aligned with the raw closed captions text in order to get automatically aligned references for the audio data. Development and test sets, manually segmented and transcribed, were also proposed inside this initial release, which is found at http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus.
The table 1 summarizes the characteristics of the data found inside the initial release of the TED-LIUM corpus.

Characteristic	Train	Dev
Number of talks	774	19
Total duration	118h 4m 48s	4h 12m 55s
- Male	81h 53m 7s	3h 13m 57s
- Female	36h 11m 41s	58m 58s
Mean duration	9m 9s	13m 18s
Number of unique speakers	666	19
Number of segments	56.8k	2k
Number of words	1.7M	47k

Table 1: The TED-LIUM corpus release 1 characteristics.

3. Selecting data for language modeling

In this section, we describe the experiments we made regarding data selection for language modeling. These experiments lead to the inclusion of well-suited selected monolingual data in our corpus. This data comes from publicly available corpora distributed within the WMT 2013 machine translation evaluation campaign[2].

[2] http://www.statmt.org/wmt13

3.1. The XenC tool

In order to select relevant data to the task of performing Automatic Speech Recognition on TED Talks, we used a data selection approach described in (Moore and Lewis, 2010), which we implemented in an open-source tool named XenC we developed and released last year (Rousseau, 2013). This approach, which is becoming commonly used in the Statistical Machine Translation field, generally helps achieving better BLEU scores and can be used both with monolingual (language model estimation) or bilingual (translation model estimation) data. In the Automatic Speech Recognition field, more than a WER or perplexity reduction (which are still usually observed), we aim at reducing the size of the training data, thus estimating smaller LMs and consequently optimizing decoding speed and disk usage.

The XenC tool filtering framework consists of the following: from an in-domain corpus and one out-of-domain corpus, we first estimate two language models. The first LM is estimated from the whole in-domain corpus. The second LM is estimated from a random subset of the out-of-domain one, with a number of tokens equal to the in-domain one. These two models are then used to compute two scores for each sentence of the out-of-domain corpus, so the difference between these scores would provide an estimation of the closeness of each out-of-domain sentence regarding the in-domain subject. Finally, the out-of-domain corpus is sorted by score, and evaluation then threshold decision are performed by computing perplexity of language models estimated from parts of various sizes of the sorted corpus. In other terms, our tool will extract cumulative parts based on a fixed step size (usually ten percent), estimate LMs on them, then compute the perplexity against a development corpus or the in-domain one.

3.2. Data selection for TED Talks

The data selected for language modeling using our XenC tool comes from corpora distributed within the WMT 2013 evaluation campaign. Each individual corpus from WMT13 is considered as out-of-domain, while our in-domain corpus consists of all the text from the transcription files of our corpus TED-LIUM release 1. The table 2 presents the characteristics of the original and selected data, in terms of number of sentences. Regarding the UN corpus, the number of selected sentences is equal to zero as this corpus seems totally out-of-domain according to our tool (very high perplexity, even when keeping only one percent).

3.3. Experiments

For these experiments, we made baseline acoustic models for the Kaldi decoder (Povey et al., 2011) by only using training data available in the TED-LIUM corpus first release, briefly described in section 2.. These models have first been trained using linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) feature transformations, then speaker adaptive training (SAT) and feature space maximum mutual information (fMMI) (Povey et al., 2008) estimation were performed.

Corpus	# Sent. original	# Sent. selected	% of original
Common Crawl	8 528 785	1 194 029	14
Europarl v7	2 268 659	180 541	8
10^9 FR-EN	22 520 400	900 530	4
News-com. v8	247 966	32 234	13
News	68 521 621	9 593 018	14
UN	14 126 730	0	0
Yandex 1M	1 000 000	350 000	35
Total	117 214 161	12 250 352	10.45

Table 2: Original and selected monolingual data characteristics.

In order to achieve a good comparison, we made three sets of quadrigram language models (4G LMs). The first one is composed of a language model estimated of the text extracted from the TED-LIUM transcriptions only a the language model which was used during the IWSLT 2011 campaign. The second set is composed of language models estimated from the whole monolingual data presented in table 2, while the third set corresponds to language models estimated from the selected data. For each of these two sets, two language models have been produced:

- the first one is estimated from a linear interpolation of individual language models for each data source,
- the second one is estimated using a single corpus where all data sources are concatenated.

The coefficients used for linear interpolations were automatically computed against the development set transcriptions from the TED-LIUM corpus. The vocabulary is the same as the one we used during the IWSLT 2011 evaluation campaign (Rousseau et al., 2011) which consists of about 150k both manually and automatically phonetized words. All language models have been estimated using the SRILM toolkit (Stolcke, 2002) with a modified Kneser-Ney discounting method and no cut-offs. The table 3 summarizes, for each language model, the number of bigrams, trigrams and quadrigrams (in millions), as well as, for 4G LMs, the perplexity obtained on the development set and the LM size on disk (in gigabytes).

We can see that data selection leads to an important reduction of both perplexity and LM size, thus improving decoding time and memory usage. We can also observe that there is almost no difference in perplexity between linear interpolation and concatenation, except a tiny reduction for the linear interpolation of all the models, compared to the concatenation. The table 4 details the different interpolation weights that were calculated for 4G LMs estimation, for both datasets (All and Selected).

This table shows that when using all data, our in-domain corpus TED-LIUM weight is very large, accounting for

LM	# 2G	# 3G	# 4G	PPL	Size
TED	0.4M	0.9M	1.1M	263	0.03G
IWSLT	35.8M	192.2M	392.7M	212	6.3G
catAll	29.1M	209.5M	573.9M	184	7.9G
linAll	29.1M	209.5M	573.9M	183	7.9G
catSel	17.2M	71M	124.1M	113	2.3G
linSel	17.2M	71M	124.1M	113	2.3G

Table 3: Characteristics of each language model: TED is TED-LIUM transcriptions data only, IWSLT is original LM from IWSLT 2011 campaign, catAll and catSel are concatenation of All and Selected data respectively, linAll and linSel are linear interpolation of All and Selected data respectively.

Corpus	linAll weights	linSel weights
TED-LIUM	0.61817	0.16573
Common Crawl	0.14552	0.36927
Europarl v7	0.00483	0.00542
10^9 FR-EN	0.01643	0.04470
News-com. v8	0.00145	0.00154
News	0.16272	0.36208
Yandex 1M	0.04882	0.05125

Table 4: Interpolation weights of the various corpora in final 4G LM estimation, for both All and Selected datasets.

more than half the total weights. We also see that Europarl and News-commentary weights are very weak and that large corpora like Common Crawl or News (with potentially many domain-related sentences) are underrepresented. The distribution of weights seems suboptimal. Conversely, when using filtered data, the weight for TED-LIUM is much lower but still of some importance, while the weights for other significant corpora like Common Crawl or News seems smoother. Weights for less significant corpora, while not very different, are at least a little stronger, meaning that the selected sentences are closer to the considered task.

3.4. Evaluation of the language models

Using each of these language models, we then performed a decoding on our development and test sets. The table 5 presents the word-error rate (WER) we obtained on each set for each language model.

In this table, we can observe that using automatically selected data with our open-source tool XenC leads to interesting word-error rate reductions. When comparing the "catSel" LM WER with the IWSLT 2011 LM WER, we report a reduction of 2.0 points (11.3% relative) on the development set and 1.0 point (5.3% relative) on the test set. When comparing the same LM WER with the "catAll" LM WER (thus indicating the impact of data selection), we can see a reduction of 0.9 point (5.4% relative) on the development set and 0.2 point (1,1% relative) on the test set. It is noticeable that the LM estimated from concatenated selected data reaches a lower WER than the LM computed from interpolated selected data. This behavior is observed because using individual corpora for selection induces inclusion of some unrelated sentences, while processing all data at once allows a better score estimation, relegating unwanted sentences altogether.

LM	Dev WER	Test WER
TED	20.9%	21.5%
IWSLT	17.7%	18.8%
catAll	16.6%	18.0%
linAll	16.6%	18.4%
catSel	**15.7%**	**17.8%**
linSel	16.0%	18.2%

Table 5: Results in terms of word-error rate (WER) obtained on TED-LIUM development and test sets for each considered language model.

4. Enhancing the corpus with new talks

In this section we describe the second enhancement we propose for the TED-LIUM corpus, which is the addition of many new TED Talks, under the form of speech segments: acoustic data and corresponding transcriptions with timings.

4.1. Selection of new segments

We extracted 753 new talks from the TED website, accounting for 158 hours of raw acoustic data, compared to the 818 talks representing 216 hours of raw acoustic data extracted for the first release of TED-LIUM. This acoustic data has been automatically segmented using our in-house tool (Meignier and Merlin, 2010) to produce 61699 speech segments. We also extracted the corresponding closed captions, representing about 1,4 million words of raw textual data.

In order to automatically align these closed captions to the acoustic data to produce proper transcriptions, we first built a deep neural network (DNN) based on state-level minimum Bayes risk (sMBR) (Kingsbury, 2009) discriminative criterion upon fMMI baseline acoustic models similar to the ones described in section 3.3.. The differences reside in an augmented dictionary with new phonetized words from the closed captions raw text and an updated 4G language model with the new vocabulary and sentences from this same text. The deep neural network has 7 layers for a total of 42.5 millions parameters and each of the 6 hidden layers has 2048 neurons. The output dimension is 10049 units and the input dimension is 440, which corresponds to an 11 frames window with 40 LDA parameters each. Weights for the network are initialized using 6 restricted Boltzmann machines (RBMs) stacked as a deep belief network (DBN). The first RBM (Gaussian-Bernoulli) is trained with a learning rate of 0.01 and the 5 following RBMs (Bernoulli-Bernoulli) are trained with a rate of 0.4. The learning rate for the DNN training is 0.00001. The segments and frames are processed randomly

during the network training with stochastic gradient descent in order to minimize cross-entropy between the training data and network output. To speed up the learning process, we used a GPU and the CUDA toolkit for our computations.

The process of segment alignment and selection can be split up into several steps:

1. first, we use the neural network and the 4G LM to decode the whole set of new segments;
2. then, we align, for each individual talk, the resulting output of the ASR system with the raw closed captions text using an algorithm based on WER minimization. The system output is considered as the reference;
3. at this step, it is advisable to remove the worst aligned talks according to their alignement WER to skip processing unwanted talks, like ones in foreign language, made of songs or spoken by non-native english speakers with very strong accents;
4. select all segments where the closed captions text perfectly matches the system's transcriptions and update the training set of segments;
5. estimate new acoustic models and neural network using the updated set of training data.

This process can be iterated several times, each time enhancing the training data with new segments. When the improvement of the system's WER starts to become too low and stability is reached, one last iteration can be done in order to select the final set of added segments to the training database. We use the same process as described above, except that we select segments where only the first and last word match between the transcriptions and the closed captions, keeping the text from the closed captions.

The table 6 summarizes the characteristics of the newly added data through each iteration.

Iteration	# of seg.	Speech hours	WER Dev fMMI	WER Dev DNN	WER Test
Baseline	56803	118h	15.7	12.4	13.5
1	71474	144h	15.5	11.4	12.4
2	75747	154h	14.0	11.2	11.9
3	77142	157h	13.9	11.0	11.8
Final	**92976**	**207h**	**13.6**	**10.4**	**11.3**

Table 6: TED-LIUM release 2 acoustic data statistics and scores by iteration. Baseline is using TED-LIUM release 1 only as training data.

After the final iteration, we end up adding more than 35000 useful segments (36173, *i.e.* 41.4% of the new talks segments) to our corpus, effectively reducing the word-error rate by 2.0 points (16.1% relative) when decoding with the neural network and 2.1 points (13.4% relative) when decoding with the fMMI models.

4.2. Characteristics of the enhanced corpus

The table 7 summarizes the characteristics of the textual and audio data of the new release of the TED-LIUM corpus. Statistics for both releases are presented, as well as the evolution between the two.

Characteristic	Corpus v1	Corpus v2	Evolution
Total duration	118h	207h	75.4%
- Male	82h	141h	72%
- Female	36h	66h	83.3%
Mean duration	9m 9s	10m 12s	11.5%
Number of unique speakers	666	1242	86.5%
Number of talks	774	1495	93.1%
Number of segments	56803	92976	63.7%
Number of words	1.7M	2.6M	52.9%

Table 7: TED-LIUM release 2 corpus audio and text characteristics.

4.3. Updating the language model

In order to propose a complete experiments, we made a second and last experiment regarding the data selection for language modeling. As described in section 3., we used our XenC tool to select data from the same corpora sets than before, but considering the text from the new training set as our in-domain corpus. We also used the updated vocabulary mentioned in section 4.1. to perform the selection and estimate the new language models. The tables 8 and 9 presents respectively the differences in data selection between the first and updated language models and the characteristics of the updated one compared to the "catSel" one described in section 3..

Corpus	% of original 1st selection	% of original 2nd selection
Common Crawl	14	9
Europarl v7	8	6
10^9 FR-EN	4	4
News-com. v8	13	9
News	14	18
UN	0	0
Yandex 1M	35	31
Total	10.45	12.34

Table 8: Differences in selection between the first and second language models.

We can see that we end up selecting 12.34% of the original corpora (18.1% more data than before) and achieving a small perplexity gain for an increase in size of only 0.3 gigabytes. The table 10 reports the performance of the system when using this updated language model for decoding.

LM	# 2G	# 3G	# 4G	PPL	Size
catSel	17.2M	71M	124.1M	113	2.3G
newCatSel	18.8M	80.6M	144.2M	111	2.6G

Table 9: Characteristics of the updated "catSel" language model compared to the one described in section 3..

LM	Dev WER	Test WER
catSel	10.4%	11.3%
newCatSel	**10.1%**	**11.1%**

Table 10: Results in terms of word-error rate (WER) obtained on TED-LIUM development and test sets for first and second version 4G language models.

In the end, our final system lead to an interesting reduction in WER of 2.3 points (18.5% relative) with our updated acoustic models and neural network on the new training set plus our updated language model.

4.4. Availability of the second release

This second release of TED-LIUM will be made available later this year, on our local website in the same place as the first one (see URL in section 2.). It will be composed of the following:

- 1 495 TED talks acoustic signal in NIST Sphere format (SPH) as training material,
- 1 495 accompanying reference transcriptions in STM format,
- 14 469 724 lines of selected text for language modeling,
- an updated phonetized dictionary of 159 849 words with variants (152 213 unique words),
- 19 TED talks in SPH format with corresponding manual transcriptions to divide into development and test sets.

5. Conclusion

In this paper, we presented the improvements we made to our previously released TED-LIUM corpus. We first showed that interesting word-error rate reductions can be obtained with language models composed of filtered data using cross-entropy difference; as well as reductions in size, thus improving the decoding time and memory usage. Then, we presented the additions made to the updated TED-LIUM corpus. We described the ASR system trained to perform the segment selection process and reported the results obtained on each iteration, as well as results using an updated language model.

6. References

Federico, M., Bentivogli, L., Paul, M., and Stüker, S. (2011). Overview of the IWSLT 2011 evaluation campaign. In *Proceedings of International Workshop on Spoken Language Translation*, pages 11–27, Décembre.

Kingsbury, B. (2009). Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009*, pages 3761–3764, April.

Meignier, S. and Merlin, T. (2010). LIUM SpkDiarization: an open source toolkit for diarization. In *Proceedings of the CMU SPUD Workshop*, Mars.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL Conference Short Papers*, pages 220–224, Juillet.

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., and Visweswariah, K. (2008). Boosted mmi for model and feature-space discriminative training. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.*, pages 4057–4060, March.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December.

Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H., and Estève, Y. (2011). LIUM's systems for the IWSLT 2011 speech translation tasks. In *Proceedings of International Workshop on Spoken Language Translation*, pages 79–85, Décembre.

Rousseau, A., Deléglise, P., and Estève, Y. (2012). TED-LIUM: an automatic speech recognition dedicated corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 125–129, Mai.

Rousseau, A. (2013). Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100(1):73–82, October.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of Interspeech*, pages 901–904, Septembre.